EXPLORING THE POTENTIAL OF MLLMS FOR ZERO-SHOT ANOMALY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing multimodal large language model (MLLM)-based methods for zero-shot anomaly segmentation (ZSAS) rely on dedicated segmentation components, leaving open whether a single MLLM can directly segment and describe anomalies. To investigate this question, we empirically examine two potential bottlenecks. First, at the vision encoder level, pretrained visual features exhibit weaker semantic relationships within regions than those of visual foundation models, potentially hindering anomaly localization that requires semantically coherent representations. Second, at the LLM level, contextual `[SEG]` representations exhibit limited diversity in their variation across input images under repetitive mask supervision, potentially restricting adaptation to different anomaly characteristics. Based on these findings, we propose **MIAS** (**M**LLM **I**tself for **A**nomaly **S**egmentation), an MLLM-only framework with two training strategies that address the respective bottlenecks. To strengthen semantic relationships in the vision encoder's features, Patch Affinity Distillation (PAD) transfers pairwise patch affinities from a frozen visual foundation teacher. To encourage more diverse image-conditioned `[SEG]` representations in the LLM, Complementary Prompt Supervision (CPS) constructs complementary localization tasks from existing anomaly masks without additional manual annotations. Both strategies operate only during training. At inference, MIAS uses a single MLLM to generate anomaly descriptions and predict masks by directly matching contextual segmentation queries with visual features, requiring neither external visual models nor dedicated mask decoders. Experiments across eight industrial and medical datasets show that MIAS achieves the best average segmentation performance among the compared methods.

1 INTRODUCTION

Zero-shot anomaly segmentation (ZSAS) aims to localize anomalous regions in unseen categories without category-specific retraining. It is relevant to industrial inspection and medical image analysis, where anomaly annotations are scarce and target appearances vary across domains. Beyond pixel-wise localization, textual descriptions can help characterize the type, location, and appearance of an anomaly. This motivates a unified model that directly produces anomaly masks and accompanying descriptions while generalizing to unseen categories.

Early ZSAS methods predominantly use CLIP-like vision-language models to compare normal and abnormal text prompts with patch embeddings Jeong et al. (2023); Chen et al. (2023); Zhou et al. (2024). Although effective for open-vocabulary anomaly scoring, these methods can struggle with subtle defects and irregular boundaries, and do not naturally generate anomaly descriptions. Multimodal large language models (MLLMs) Liu et al. (2023); Chen et al. (2024); Bai et al. (2025) offer strong generalization capabilities and a language interface for combining visual localization with textual outputs, making them a promising foundation for ZSAS.

MLLMs have shown promising capabilities in semantic and referring segmentation. Models such as LISA Lai et al. (2024) and GlaMM Rasheed et al. (2024) connect contextual segmentation tokens to external mask decoders, while recent works such as UFO Tang et al. (2026) and SELF1E Zhang et al. (2026) demonstrate decoder-free mask prediction for general segmentation. Despite these advances, the potential of MLLMs for ZSAS remains underexplored. AG-VAS Qu et al. (2026) extends MLLM-based reasoning to zero-shot visual anomaly segmentation but still relies on a dedi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

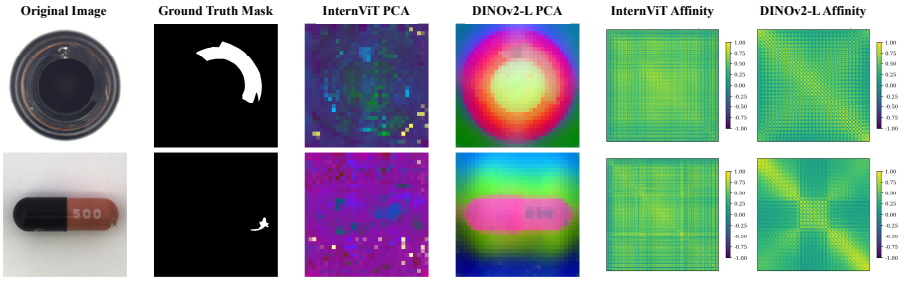


Figure 1: Visual representation diagnosis. Final-layer PCA projections and patch affinity matrices show weaker spatial organization and noisier patch relationships in pretrained InternViT than in DINOv2-L, motivating structural supervision for direct mask prediction.

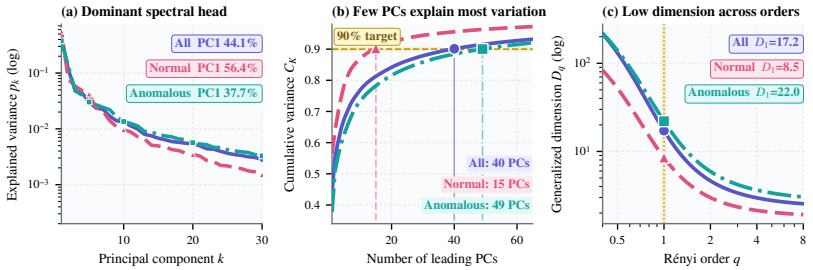


Figure 2: Query representation diagnosis under repeated, unchanged supervision. With a fixed inference prompt, (a) the centered projected-query spectrum has a dominant head, (b) few components explain most sample-dependent variation, and (c) generalized dimensions remain low across Rényi orders. Results cover all 1, 725 MVTec-AD test images and their normal and anomalous subsets.

cated pixel encoder and mask decoder. This leaves open our question: *can a single MLLM segment and describe anomalies in unseen categories without external visual models or mask decoders?*

To investigate this question, we adopt the embedding-as-mask paradigm, where a contextual segmentation embedding is matched with visual tokens to predict a mask. However, strong multimodal understanding does not automatically translate into precise anomaly masks. Accurate localization depends on two complementary representations: visual features that preserve semantic relationships within regions and a segmentation query that captures image-specific localization cues. Our preliminary analyses of the vision encoder and the LLM reveal two potential representation bottlenecks.

First, at the vision encoder level, pretrained visual features exhibit weaker semantic relationships within regions than those of visual foundation models. Unlike complete semantic objects, anomalies may consist of sparse, irregular, or texture-level deviations. Their localization requires associating patches within meaningful regions while distinguishing anomalies from surrounding normal content. On the same input images, pretrained InternViT features exhibit fragmented spatial organization and noisy affinities, whereas visual foundation models such as DINOv2 Oquab et al. (2023) produce more coherent region-level structure (Figure 1). Such weaker semantic organization may hinder coherent anomaly localization through direct query–patch matching.

Second, at the LLM level, contextual `[SEG]` representations exhibit limited diversity in their variation across images. In a control trained with repeated, unchanged image–instruction–mask pairs, normalized projected `[SEG]` queries concentrate in few principal directions under a fixed inference instruction (Figure 2). This concentration reflects limited diversity in image-dependent query directions after removing magnitude differences and the dataset-level component. Repeated supervision of the same localization objective may therefore provide insufficient constraints for learning diverse, image-conditioned queries, potentially restricting adaptation to different anomaly characteristics.

Based on these findings, we propose **MIAS** (**M**LLM **I**tself for **A**nomaly **S**egmentation), an MLLM-only framework with two training strategies addressing the respective bottlenecks. Patch Affinity Distillation (PAD) aligns pairwise patch affinities with those of a frozen visual foundation teacher, transferring region-level relational structure without channel-wise feature regression to strengthen semantic relationships in visual features. Complementary Prompt Supervision (CPS) pairs complementary localization objectives derived from existing anomaly masks with region-specific instruc-

tions. This enriches supervision to encourage more diverse image-conditioned [SEG] representations without additional images or manual annotations.

Both strategies operate only during training. At inference, the teacher is discarded, and complementary training prompts are replaced by a single task instruction. MIAS directly matches the MLLM's own visual features with contextual segmentation queries to predict masks alongside generated anomaly descriptions, requiring neither external visual models nor dedicated mask decoders.

With a 2B MLLM and limited task-specific training data, MIAS achieves the best average segmentation performance among the compared methods across eight industrial and medical datasets. Qualitative examples also demonstrate joint anomaly description and localization. These results support targeted supervision of visual and contextual query representations as an effective route to MLLM-only anomaly segmentation.

The main contributions are summarized as follows:

- We investigate MLLM-only ZSAS and identify two potential representation bottlenecks: weak semantic relationships within regions at the vision encoder level and limited diversity of image-dependent contextual [SEG] representations at the LLM level.
- We propose MIAS with two targeted training strategies: PAD strengthens semantic relationships in visual features through patch-affinity transfer, while CPS enriches localization supervision to encourage more diverse image-conditioned segmentation queries.
- We demonstrate the best average zero-shot segmentation performance among the compared methods across eight industrial and medical datasets, together with qualitative examples of joint anomaly description and segmentation, using a single MLLM without external visual models or dedicated mask decoders at inference.

2 Related Work

2.1 Zero-Shot Anomaly Segmentation

Zero-shot anomaly segmentation has been studied extensively in the CLIP Radford et al. (2021) era. Methods such as WinCLIP Jeong et al. (2023), April-GAN Chen et al. (2023) and AnomalyCLIP Zhou et al. (2024) build anomaly scores from prompt-based alignment between image patches and textual concepts. Later methods such as AdaCLIP Cao et al. (2024), AA-CLIP Ma et al. (2025) and Bayes-PFL Qu et al. (2025) introduce prompt learning or Bayesian prompt flows to improve fine-grained discrimination. These methods are appealing because they require no defect-specific training categories; however, their performance is often limited when anomalies have irregular shapes or weak semantic cues, and the architecture of CLIP-style models inherently restricts their ability to perform interactive anomaly reasoning and instruction-driven anomaly understanding.

2.2 MLLM-Based Segmentation

LISA Lai et al. (2024) and GlaMM Rasheed et al. (2024) connect contextual segmentation tokens to specialist mask decoders. UFO Tang et al. (2026) and SELF1E Zhang et al. (2026) instead generate masks by directly matching segmentation embeddings with high-resolution visual tokens, but focus primarily on semantic or referring segmentation. AG-VAS Qu et al. (2026) extends MLLMs to anomaly segmentation using a heavier architecture and additional supervision. In contrast, our work conditions the two representations used by decoder-free mask prediction: PAD transfers pairwise spatial structure without channel-wise feature alignment, while CPS diversifies contextual segmentation queries by expanding the supervision space of existing annotations.

3 Preliminary Analysis

For embedding-as-mask prediction, let $\mathbf{F}_i \in \mathbb{R}^{P\times d}$ denote the dense prediction features for image $\mathbf{x}_i$ and instruction $\mathbf{q}_i$, and $\mathbf{z}_i \in \mathbb{R}^d$ the projected contextual state associated with [SEG]. Unlike the

shared token embedding, $\mathbf{z}_i$ depends on both the image and instruction. The mask probabilities are

$$\widehat{\mathbf{m}}_i = \sigma\left(\frac{\mathbf{F}_i \mathbf{z}_i}{\sqrt{d}}\right), \tag{1}$$

where P counts spatial positions and σ is the element-wise sigmoid. This direct interface motivates examining both spatial organization of visual features and sample-dependent variation of the query.

Structural organization of the visual field. For two patch features $\mathbf{f}_p$ and $\mathbf{f}_q$ from the same image, matched with a fixed query $\mathbf{z}$, the Cauchy–Schwarz inequality and $\sup_t \sigma'(t) = 1/4$ give

$$|\widehat{m}_p - \widehat{m}_q| \le \frac{|\mathbf{z}^\top(\mathbf{f}_p - \mathbf{f}_q)|}{4\sqrt{d}} \le \frac{\|\mathbf{z}\|_2}{4\sqrt{d}}\|\mathbf{f}_p - \mathbf{f}_q\|_2. \tag{2}$$

At a fixed query norm, nearby features yield similar responses, motivating within-region consistency while preserving distinctions across anomaly boundaries. Distant features, however, need not produce distinct responses, which also depend on query alignment.

Pairwise affinity provides a way to characterize these relationships. For unit-normalized features, $\|\widetilde{\mathbf{f}}_p - \widetilde{\mathbf{f}}_q\|_2^2 = 2(1 - A_{pq})$, where $A_{pq} = \widetilde{\mathbf{f}}_p^\top \widetilde{\mathbf{f}}_q$. Thus, normalized affinities encode pairwise distances on the unit sphere without requiring channel-wise correspondence between feature spaces.

The contrast between fragmented InternViT features and the more coherent region-level organization of DINOv2-L in Figure 1 motivates improving the spatial coherence of MLLM visual features while preserving distinctions needed for anomaly localization.

Concentration of contextual query variation. To examine image-dependent query directions under a fixed inference instruction, we normalize each query as $\mathbf{u}_i = \mathbf{z}_i/\|\mathbf{z}_i\|_2$ and center it as $\boldsymbol{\delta}_i = \mathbf{u}_i - \frac{1}{N}\sum_j \mathbf{u}_j$ over N evaluated images. Normalization removes query-magnitude differences, while centering removes the dataset-level common component, isolating directional variation across images. Let $\mathbf{U}_K$ contain the leading K orthonormal eigenvectors of the sample covariance of $\boldsymbol{\delta}_i$, with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots$. Summing the squared orthogonal-projection residuals yields

$$\frac{1}{N-1}\sum_{i=1}^{N} \left\|\boldsymbol{\delta}_i - \mathbf{U}_K \mathbf{U}_K^\top \boldsymbol{\delta}_i\right\|_2^2 = \sum_{k>K} \lambda_k. \tag{3}$$

This identity equates the variation lost by retaining only the leading K directions with the spectral tail energy. A small tail relative to total variance indicates that most directional variation lies in a low-dimensional subspace. We use effective rank to quantify the spread of variance across principal directions, with formal definitions of the spectral metrics provided in Appendix A.1.

In the control with unchanged copies of anomalous training samples, projected queries exhibit a dominant spectral head (Figure 2): across all 1,725 MVTec-AD test images, PC1 explains 44.1% of centered variance and the effective rank is 17.20. This concentration suggests that repeated supervision of one localization objective may provide limited incentives for diverse contextual query directions, motivating richer localization supervision.

However, spectral concentration alone does not establish poor segmentation, because visual features also vary across images. Its relevance to localization therefore requires both controlled representation analysis and segmentation evaluation.

4 METHOD

Given an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ and a task instruction $\mathbf{q}$, MIAS generates a textual response and uses the contextual state that predicts `[SEG]` to construct a segmentation query. This query is matched with dense visual features to produce a mask without a specialist decoder. Figure 3 summarizes the framework, with PAD and CPS supervising visual and query representations during training.

4.1 LIGHTWEIGHT EMBEDDING-AS-MASK SEGMENTATION

Let $\mathbf{F}_{\text{img}}^{\text{cmp}} \in \mathbb{R}^{N_1\times d}$ denote the compressed visual tokens supplied to the MLLM, and $\mathbf{F}_{\text{img}}^{\text{hr}} \in \mathbb{R}^{N_0\times d}$ a high-resolution branch retained before spatial compression, where $N_0 > N_1$. The compressed tokens support efficient multimodal processing, while the high-resolution branch preserves spatial detail.

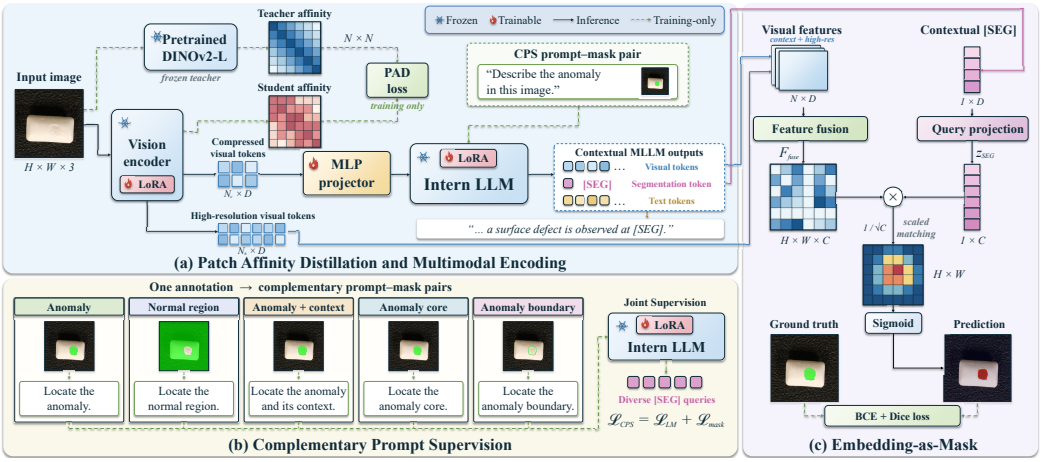


Figure 3: Overview of MIAS. PAD distills pairwise patch relationships from a frozen DINOv2 teacher, while CPS diversifies contextual segmentation queries through complementary region–instruction pairs derived from existing masks.

Let $\mathbf{F}_{\text{img}} \in \mathbb{R}^{N_1 \times d}$ denote the corresponding visual tokens after MLLM processing. We extract their residual update and fuse it with the high-resolution branch:

$$\mathbf{F}_{\text{res}} = \mathbf{F}_{\text{img}} - \mathbf{F}_{\text{img}}^{\text{cmp}}, \qquad \mathbf{F}_{\text{img}}^{\text{fuse}} = \mathbf{F}_{\text{img}}^{\text{hr}} + \mathcal{U}(\mathbf{F}_{\text{res}}), \tag{4}$$

where $\mathcal{U}$ spatially upsamples the residual. This combines retained spatial detail with the visual-feature update introduced by the MLLM.

A learned projection maps the contextual state $\mathbf{h}_i$ that predicts `[SEG]` to $\mathbf{z}_i = g_\theta(\mathbf{h}_i) \in \mathbb{R}^d$. The main mask follows Equation 1 with $\mathbf{F}_i = \mathbf{F}_{\text{img},i}^{\text{fuse}}$, and is reshaped and upsampled to the input resolution. We denote this output by $\widehat{\mathbf{m}}_{i,\text{main}}$ and the auxiliary lower-resolution prediction used during training by $\widehat{\mathbf{m}}_{i,\text{low}}$. Both receive mask supervision, as specified in Section 4.4.

4.2 Patch Affinity Distillation

Let $\mathbf{V}^l = \{\mathbf{v}_i^l\}_{i=1}^{N_s}$ and $\mathbf{D}^l = \{\mathbf{d}_i^l\}_{i=1}^{N_t}$ denote student and frozen DINOv2 teacher patch tokens at layer l. To accommodate different spatial resolutions, we reshape the student and teacher tokens into grids and bilinearly resize them to a common $r \times r$ resolution, denoted by $\bar{\mathbf{V}}_l = \mathcal{R}(\mathbf{V}_l, r)$ and $\bar{\mathbf{D}}_l = \mathcal{R}(\mathbf{D}_l, r)$, where $\mathcal{R}$ denotes bilinear resizing. Each resized token is normalized as $\widetilde{\mathbf{v}}_i^l = \bar{\mathbf{v}}_i^l / \|\bar{\mathbf{v}}_i^l\|_2$, and analogously for $\widetilde{\mathbf{d}}_i^l$. Their pairwise affinities are

$$\mathbf{A}_s^l(i,j) = \left\langle \widetilde{\mathbf{v}}_i^l, \widetilde{\mathbf{v}}_j^l \right\rangle, \qquad \mathbf{A}_t^l(i,j) = \left\langle \widetilde{\mathbf{d}}_i^l, \widetilde{\mathbf{d}}_j^l \right\rangle. \tag{5}$$

For each patch, we exclude self-similarity and convert the off-diagonal affinities into probability distributions $\mathbf{P}_s^l(i) = \mathcal{N}(\mathbf{A}_s^l(i, \neg i))$ and $\mathbf{P}_t^l(i) = \mathcal{N}(\mathbf{A}_t^l(i, \neg i))$, where $\mathcal{N}$ denotes affinity-row probability normalization. With $n = r^2$ patches and selected layers $\mathcal{L}$, PAD minimizes the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{PAD}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{1}{n} \sum_{i=1}^{n} D_{\text{KL}} \left(\mathbf{P}_t^l(i) \,\|\, \mathbf{P}_s^l(i) \right). \tag{6}$$

PAD transfers relational structure without requiring channel-wise feature correspondence or a learned channel adapter. Inner-product affinities remain invariant to independent orthogonal transformations of the teacher and student feature spaces. The teacher forward pass and $O(n^2)$ affinity computation are confined to training.

4.3 Complementary Prompt Supervision

CPS enriches localization supervision by associating the same annotated image with complementary region requests. For an anomalous image $\mathbf{x}_i$ and mask $\mathbf{m}_i$, it constructs five objectives: the complete

anomaly, the normal region, a dilated context region, an eroded anomaly core, and a narrow boundary band. Each objective v yields an instruction–target pair $(\mathbf{q}_i^{(v)}, \mathbf{m}_i^{(v)})$ through deterministic mask transformations, requiring no additional source images or manual annotations.

If erosion removes a small defect entirely, the original mask is retained. Normal images keep their empty anomaly target and receive no complementary region objectives. The instruction $\mathbf{q}_i^{(v)}$ conditions the contextual state $\mathbf{h}_i^{(v)}$ and projected query $\mathbf{z}_i^{(v)} = g_\theta(\mathbf{h}_i^{(v)})$. Both predictions $\widehat{\mathbf{m}}_{i,b}^{(v)}$, $b \in \{\text{main}, \text{low}\}$, are supervised by the corresponding target $\mathbf{m}_i^{(v)}$.

These objectives require the model to support distinct region requests for the same image, strengthening instruction-conditioned localization supervision. The complementary objectives are training-only. For quantitative segmentation evaluation, each image receives the fixed instruction "Please segment all anomalous regions in this image." and uses a single `[SEG]` query. Joint description and segmentation use a corresponding task instruction through the same prediction interface.

4.4 Training Objective

Following Anomagic Jiang et al. (2026), we use AnomVerse to augment anomaly image–mask pairs with MLLM-generated descriptions, providing language supervision without manual annotation. For image i and localization objective v, the task loss combines autoregressive cross-entropy with binary cross-entropy (BCE) and Dice losses for both mask branches:

$$\mathcal{L}_{i,v}^{\text{CPS}} = \lambda_{\text{text}}\mathcal{L}_{\text{CE}} + \sum_{b\in\{\text{main,low}\}} \left[\lambda_{\text{bce}}\mathcal{L}_{\text{BCE}}\left(\widehat{\mathbf{m}}_{i,b}^{(v)}, \mathbf{m}_i^{(v)}\right) + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}\left(\widehat{\mathbf{m}}_{i,b}^{(v)}, \mathbf{m}_i^{(v)}\right)\right]. \tag{7}$$

We use $\lambda_{\text{text}} = 1.0$, $\lambda_{\text{bce}} = 2.0$, and $\lambda_{\text{dice}} = 0.5$.

For a minibatch $\mathcal{B}$ of image–supervision pairs, the total objective is

$$\mathcal{L}_{\text{total}} = \frac{1}{|\mathcal{B}|}\sum_{(i,v)\in\mathcal{B}} \mathcal{L}_{i,v}^{\text{CPS}} + \lambda_{\text{pad}}\mathcal{L}_{\text{PAD}}. \tag{8}$$

PAD is computed once per image, preventing multiple localization objectives from implicitly reweighting affinity supervision.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on eight datasets spanning industrial inspection and medical imaging. The industrial group comprises MVTec-AD Bergmann et al. (2019), VisA Zou et al. (2022), MPDD Jezek et al. (2021) and RSDD Niu et al. (2020); the medical group comprises BrainMRI and Liver CT from BMAD Bao et al. (2024), CVC-ColonDB Tajbakhsh et al. (2015), and CVC-ClinicDB Bernal et al. (2015). Unless evaluating VisA, the model is trained on 2,162 unique VisA images and tested directly on every target dataset without category-specific fine-tuning. To prevent train–test overlap, VisA is evaluated with a model trained on MVTec-AD.

Implementation Details. We use InternVL3-2B Zhu et al. (2025b) as the backbone and DINOv2-ViT-L as the teacher, with an input resolution of 448×448. Training uses AdamW with a peak learning rate of 5×10^{-5}, cosine decay with warmup, an effective batch size of 64, and $\lambda_{\text{pad}} = 1.0$. We apply rank-8 LoRA Hu et al. (2022) to both the language model and vision transformer. PAD aligns only the final-layer features (block 24) of DINOv2 and InternViT. All experiments are conducted on two NVIDIA GeForce RTX 5090 GPUs.

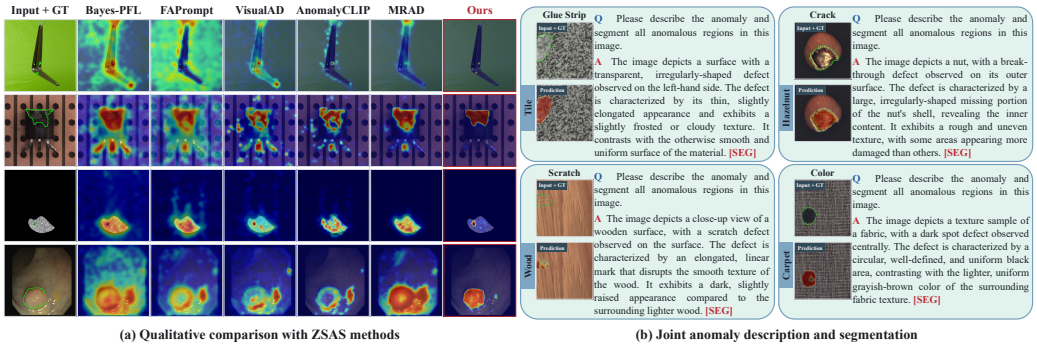
Evaluation Metrics. We report pixel-level area under the receiver operating characteristic curve (AUROC), average precision (AP), and maximum F1 over all thresholds (F1-max). Dataset averages are macro-averages over the eight datasets in Table 1.

5.2 Main Results

Table 1 compares MIAS with AnomalyCLIP Zhou et al. (2024), Bayes-PFL Qu et al. (2025), FAPrompt Zhu et al. (2025a), VisualAD Hou et al. (2026), and MRAD Xu et al. (2026) across

Table 1: Zero-shot anomaly segmentation performance across industrial and medical AD datasets. The best results are shown in red, while the second-best results are shown in blue. Ties at the displayed precision are resolved using values rounded to two decimal places. Metrics are pixel-level AUROC/F1-max/AP, and inference latency is reported in milliseconds per image (ms/img).

Domain	Dataset	AnomalyCLIP (ICLR 2024)	Bayes-PFL (CVPR 2025)	FAPrompt (ICCV 2025)	VisualAD (CVPR 2026)	MRAD (ICLR 2026)	**MIAS (Ours)**
Industrial	MVTec-AD	(91.1,38.4,34.2)	(92.1,49.2,48.1)	(89.4,40.2,36.5)	(90.9,43.1,40.3)	(92.9,45.1,43.2)	(92.9,51.3,49.9)
	VisA	(95.5,28.5,21.3)	(96.0,34.3,29.1)	(95.8,28.2,21.9)	(95.6,34.2,27.8)	(95.7,31.4,24.4)	(95.1,35.5,30.3)
	MPDD	(95.6,32.3,26.7)	(97.8,35.0,33.0)	(93.1,35.0,30.3)	(97.1,39.9,35.1)	(97.7,40.0,34.9)	(94.3,42.1,38.3)
	RSDD	(98.9,30.9,20.4)	(99.6,42.8,34.3)	(98.4,21.4,13.4)	(99.6,43.0,35.3)	(98.6,14.3,9.4)	(99.6,56.9,54.8)
Medical	BrainMRI	(96.4,53.8,56.6)	(95.9,45.7,43.6)	(96.2,52.7,56.7)	(95.6,47.6,45.3)	(96.4,51.3,54.6)	(96.6,52.0,58.0)
	Liver CT	(93.9,6.3,3.2)	(96.6,11.9,6.1)	(94.5,7.0,3.3)	(92.7,12.0,5.7)	(96.5,10.5,5.3)	(97.6,16.9,12.0)
	ColonDB	(83.1,39.9,33.9)	(83.4,39.6,31.2)	(82.0,39.1,34.1)	(77.0,35.5,27.3)	(84.7,41.9,35.1)	(84.9,42.8,40.9)
	ClinicDB	(84.9,45.1,41.5)	(88.5,51.7,50.1)	(83.2,46.1,43.5)	(84.3,45.4,37.8)	(87.3,50.2,49.4)	(89.1,60.9,63.8)
Average		(92.4,34.4,29.7)	(93.7,38.8,34.4)	(91.6,33.7,30.0)	(91.6,37.6,31.8)	(93.7,35.6,32.0)	(93.8,44.8,43.5)
Speed (ms/img)		312.5	190.1	51.3	43.3	272.0	39.1



(a) Qualitative comparison with ZSAS methods (b) Joint anomaly description and segmentation

Figure 4: Qualitative results of MIAS. Green contours mark ground-truth anomaly boundaries; warmer colors indicate higher anomaly scores.

eight datasets. MIAS achieves the best average AUROC, F1-max, and AP, exceeding the strongest competing averages by 0.1, 6.0, and 9.1 points, respectively. The larger F1-max and AP gains suggest that improved patch and query representations primarily benefit foreground localization, with a smaller gain in overall pixel-level ranking. MIAS achieves the best F1-max and AP on all four industrial datasets and leads in 11 of 12 medical metric comparisons, with BrainMRI F1-max as the only exception, where AnomalyCLIP performs best. Achieved without medical training images, these results demonstrate generalization from industrial inspection to medical anomaly segmentation.

Figure 4(a) compares representative industrial and medical examples. MIAS produces more concentrated anomaly responses with less activation in non-target regions. It localizes small bracket defects in MPDD and the damaged transistor region in MVTec-AD while suppressing responses on intact structures and surrounding textures. On Liver CT and ClinicDB, its predictions focus more closely on the annotated lesion and polyp, respectively. Figure 4(b) illustrates joint anomaly description and segmentation through the shared `[SEG]` interface. Examples cover glue strips, cracks, scratches, and color anomalies, with descriptions capturing shape, texture and contrast with surrounding materials, complementing pixel-level localization with textual characterization.

5.3 Evolution of Patch Representations under PAD

To examine how PAD shapes the MLLM vision encoder's representations, Figure 5 visualizes patch features of InternViT at different training stages using Principal Component Analysis (PCA), projecting them into RGB space through the first three principal components. From fragmented representations at Step 0, coarse object structure emerges by Step 200, followed by greater within-region consistency and reduced background fragmentation

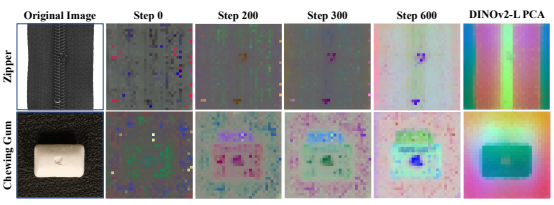


Figure 5: Evolution of InternViT features under PAD.

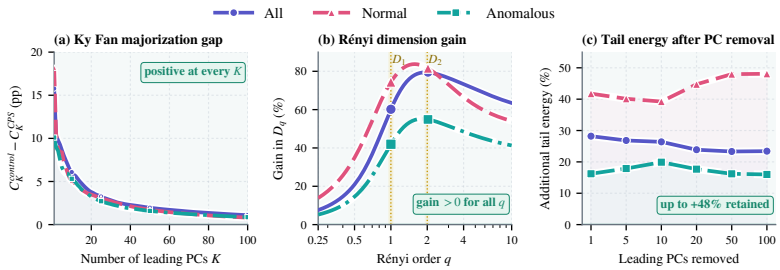


Figure 6: Spectrum-wide effects of CPS on projected queries. (a) Cumulative variance gaps, (b) generalized dimensions, and (c) normalized tail energy indicate reduced spectral concentration across all, normal, and anomalous images.

at Step 300. By Step 600, object structures and region boundaries become clearer, consistent with improved spatial coherence under PAD. By constraining relationships among patches rather than directly matching feature values, PAD encourages consistent representations within regions while preserving distinctions across boundaries.

5.4 Full-Set Analysis of Contextual [SEG] Query Collapse

Controlled protocol. We compare CPS with an identity-copy control that adds four unchanged copies of each anomalous training sample, matching the number of optimization instances and the training schedule. Only the additional supervision differs. Both checkpoints are evaluated with the same fixed segmentation-only prompt. We extract the 1,536-dimensional LLM states $\mathbf{h}_i$ and projected queries $\mathbf{z}_i$ from all 1,725 MVTec-AD test images (467 normal and 1,258 anomalous), and analyze their normalized, centered representations using the complete matched set.

Query diversification under a fixed prompt. Table 2 directly addresses the query concentration identified in our preliminary analysis. For projected queries, CPS increases effective rank from 17.20 to 27.55 and participation rank from 4.64 to 8.33, while reducing PC1's variance share from 44.1% to 28.3% and increasing PC90 from 40 to 53. CPS also increases LLM-state effective rank from 15.71 to 25.47 and reduces PC1's share from 45.2% to 29.1%, showing that the diversification emerges before query projection and under the same fixed inference prompt. The effect also extends to normal images, whose effective rank increases from 8.52 to 14.85.

Table 2: Projected [SEG] query geometry on the MVTec-AD test set. "Copy" denotes unchanged supervision. Formal definitions of the spectral metrics are provided in Appendix A.1.

Split	Supervision	$r_{\text{eff}}\uparrow$	$r_{\text{part}}\uparrow$	$p_1\downarrow$	PC90$\uparrow$
All	Copy	17.20	4.64	0.441	40
	CPS	**27.55**	**8.33**	**0.283**	**53**
Normal	Copy	8.52	3.00	0.564	15
	CPS	**14.85**	**5.44**	**0.383**	**24**
Anomalous	Copy	22.04	5.93	0.377	49
	CPS	**31.29**	**9.19**	**0.276**	**60**

Spectrum-wide evidence. Figure 6 shows that CPS reduces spectral concentration beyond effective rank. In panel (a), $C_K^{\text{control}} - C_K^{\text{CPS}}$ remains positive throughout the plotted range of $K \le 100$ for all three splits, indicating that CPS consistently shifts variance away from the leading K components within the examined range. Panel (b) shows higher generalized dimensions across $q \in [0.25, 10]$, with gains of 7.6–79.5% on the full set and consistent increases on both subsets. Panel (c) further shows 23–28%, 39–48%, and 16–20% more normalized tail energy after removing leading components for all, normal, and anomalous images, respectively. Together, these results indicate broader variance redistribution beyond the dominant principal component. Appendix A.2 further shows that CPS increases within-category effective rank in 13 of 15 MVTec-AD categories.

5.5 Ablation Studies and Discussion

We ablate both proposed mechanisms, structural supervision objective, MLLM backbone, and inference latency on MVTec-AD. Additional sensitivity studies on the alignment depth and teacher backbone are provided in Appendices A.4 and A.5. All variants use the same data split.

Effect of PAD and CPS. Table 3 evaluates the individual and combined effects of PAD and CPS. Relative to the plain embedding-as-mask baseline, CPS improves AUROC/F1-max/AP by 1.4/1.8/3.7 points, while PAD improves them by 1.6/3.5/4.8 points. Combining both yields the strongest

Table 3: Ablation of PAD and CPS. Best results in **bold**.

PAD	CPS	AUROC	F1-max	AP
×	×	90.0	46.7	41.7
✓	×	91.6	50.2	46.5
×	✓	91.4	48.5	45.4
✓	✓	**92.9**	**51.3**	**49.9**

Table 6: Comparison of MIAS+ and AG-VAS. Performance is reported as dataset-level mean F1-max and AP on common benchmarks; AG-VAS results are taken from its original paper.

Method	MLLM Params.	Training Data	MVTec-AD		ColonDB		ClinicDB	
			F1-max	AP	F1-max	AP	F1-max	AP
MIAS+	**2B**	**8.2K**	52.58	**52.71**	**74.77**	**81.25**	**85.54**	**91.81**
AG-VAS	7B	40K	**52.70**	51.00	66.20	70.70	79.20	86.60

scores, supporting complementary rather than redundant effects: PAD regularizes the patch field, whereas CPS conditions the query.

Choice of Structural Supervision Objective. Table 4 compares direct feature alignment with patch-affinity distillation (PAD) under different objectives. Direct L2 regression yields substantially lower performance compared with affinity-based L2 alignment, which supports the benefit of transferring relational patch geometry rather than directly matching feature representations. We further compare L2, Huber, cosine, Jensen–Shannon (JS), and KL objectives within PAD. Among them, KL achieves the best F1-max and AP while tying L2 for the highest AUROC. Cosine and JS are also competitive, suggesting that PAD is relatively robust to the specific alignment objective, while distribution-based affinity matching provides the strongest overall localization performance.

Table 4: Comparison of different structural supervision objectives. Best results in **bold**.

Alignment	Loss	AUROC	F1-max	AP
Direct	L2	88.8	43.1	38.2
PAD	L2	**92.9**	50.4	48.0
PAD	Huber	92.2	50.2	47.4
PAD	Cosine	92.6	51.0	48.9
PAD	JS	92.5	51.1	49.1
PAD	KL	**92.9**	**51.3**	**49.9**

Influence of the MLLM Backbone. Table 5 compares three MLLM backbones under the same experimental setting. InternVL3-2B achieves the highest F1-max and AP, exceeding InternVL3.5-2B by 2.4 and 3.1 points, respectively. InternVL3.5-2B obtains the highest AUROC, slightly above InternVL3-2B. Qwen3.5-2B yields the lowest AUROC among the three backbones, but surpasses InternVL3.5-2B on F1-max and AP, trailing only InternVL3-2B. Overall, all three backbones achieve competitive performance on MVTec-AD, suggesting that the proposed method remains effective across different MLLM backbones. A further PAD/CPS ablation with Qwen3.5-2B is provided in Appendix A.3, where the two training strategies remain effective.

Table 5: Ablation of different MLLM backbones. Best results in **bold**.

MLLM Backbone	AUROC	F1-max	AP
InternVL3.5-2B	**93.1**	48.9	46.8
InternVL3-2B	92.9	**51.3**	**49.9**
Qwen3.5-2B	92.8	50.4	49.5

Inference Latency. We compare inference latency on MVTec-AD in Table 1. Under the same hardware environment, MIAS achieves the lowest latency of 39.1 ms per image, compared with 43.3 ms for VisualAD, the next fastest method. Together with its segmentation results, this demonstrates efficient anomaly localization with a compact 2B MLLM and no dedicated mask decoder.

Discussion with AG-VAS. Throughout the paper, MIAS denotes the model trained with the standard training set in Section 5.1. We further consider an extended-data variant, MIAS+, which keeps the same MLLM backbone, decoder-free architecture, PAD and CPS objectives, and training procedure while adding VisA, 5K randomly selected RealIAD Wang et al. (2024) images, and Kvasir-SEG Jha et al. (2019) to the training data. Thus, MIAS+ differs from MIAS only in training data. Table 6 compares MIAS+ with AG-VAS Qu et al. (2026) on common benchmarks. Since AG-VAS does not release model weights for direct reproduction, we use its reported results. Despite using a 2B MLLM and decoder-free architecture, MIAS+ achieves higher AP on all three datasets and substantially higher F1-max on ColonDB and ClinicDB. On MVTec-AD, the methods have comparable F1-max, while MIAS+ achieves higher AP. These results show that MIAS+ provides competitive cross-domain anomaly localization with a smaller MLLM and lighter architecture, while requiring substantially less anomaly-specific training data.

6 Conclusion

We investigate whether an MLLM can directly segment and describe anomalies in unseen categories. We identify two bottlenecks: weak within-region visual semantics and limited diversity of image-conditioned `[SEG]` representations. Our MIAS addresses them with Patch Affinity Distillation and Complementary Prompt Supervision during training. At inference, it matches contextual queries with MLLM visual features without external models or mask decoders. Across eight datasets, MIAS achieves strong average segmentation and joint anomaly description and localization. Future work could explore more adaptive complementary objectives for richer `[SEG]` queries.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic textual descriptions used as language supervision. Specifically, following the data augmentation procedure of Anomagic Jiang et al. (2026), we used the Doubao-Seed-1.6-thinking model to generate descriptive texts for anomalous images in Section 4.4. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, assisting with proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, cleaning or reformatting datasets, or conducting qualitative or thematic analysis. Additionally, we used GPT-5.6-Sol to aid and polish the writing of this paper. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide reproducibility details throughout the paper and appendix. Sections 4.1–4.4 describe the model architecture, Patch Affinity Distillation (PAD), Complementary Prompt Supervision (CPS), and the complete training objective. Section 5.1 reports the datasets, train–test protocol, model configuration, optimization settings, LoRA configuration, hardware, and evaluation metrics. Appendix A.1 provides the formal definitions of the spectral metrics used in the query representation analysis, while Appendix A.2 examines query diversity at the category level on MVTec-AD. Appendix A.3 reports the PAD/CPS ablation with Qwen3.5-2B, and Appendices A.4 and A.5 provide sensitivity studies on the teacher–student alignment layer and the affinity teacher backbone, respectively. Together, these sections document the methodology, experimental protocol, and supplementary analyses needed to reproduce the reported results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *CVPR*, 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graphics*, 2015.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *ECCV*, 2024.

Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. *arXiv preprint arXiv:2603.07952*, 2026.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *ICUMT*, 2021.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM*, 2019.

Yuxin Jiang, Wei Luo, Hui Zhang, Qiyu Chen, Haiming Yao, Weiming Shen, and Yunkang Cao. Anomagic: Crossmodal prompt-driven zero-shot anomaly generation. In *AAAI*, 2026.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *CVPR*, 2025.

Menghui Niu, Kechen Song, Liming Huang, Qi Wang, Yunhui Yan, and Qinggang Meng. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Trans. Ind. Inform.*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *CVPR*, 2025.

Zhen Qu, Xian Tao, Xiaoyi Bao, Dingrong Wang, ShiChen Qu, Zhengtao Zhang, and Xingang Wang. Ag-vas: Anchor-guided zero-shot visual anomaly segmentation with large multimodal models. *arXiv preprint arXiv:2603.01305*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.

Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging*, 2015.

Hao Tang, Chen-Wei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. *NeurIPS*, 2026.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, 2024.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. Mrad: Zero-shot anomaly detection with memory-driven retrieval. *arXiv preprint arXiv:2602.00522*, 2026.

Anqi Zhang, Xiaokang Ji, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Rethinking mllm itself as a segmenter with a single segmentation token. *arXiv preprint arXiv:2603.19026*, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024.

Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *ICCV*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

A Appendix

A.1 Supplementary Definitions for Query Spectral Analysis

For a fixed inference instruction, we normalize each projected query and remove the dataset mean:

$$\mathbf{u}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}, \qquad \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i, \qquad \boldsymbol{\delta}_i = \mathbf{u}_i - \boldsymbol{\mu}. \tag{9}$$

The sample covariance and its normalized eigenvalue spectrum are

$$\boldsymbol{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{\delta}_i\boldsymbol{\delta}_i^{\top}, \qquad p_k = \frac{\lambda_k}{\sum_j \lambda_j}, \tag{10}$$

where $\lambda_1 \geq \cdots \geq \lambda_R > 0$ are the non-zero eigenvalues. The effective rank is

$$r_{\text{eff}} = \exp\left(-\sum_{k=1}^{R} p_k \log p_k\right). \tag{11}$$

The same procedure is applied separately to LLM states $\mathbf{h}_i$ and projected queries $\mathbf{z}_i$, and independently to the full test set and its normal and anomalous subsets.

Let $p_1 \geq \cdots \geq p_R > 0$ denote the normalized covariance eigenvalues. In addition to the effective rank in Equation 11, we report the participation rank

$$r_{\text{part}} = \frac{1}{\sum_{k=1}^{R} p_k^2}, \tag{12}$$

the leading-component share p_1, and the number of components required to explain at least 90% of the variance:

$$\text{PC90} = \min\left\{K : \sum_{k=1}^{K} p_k \geq 0.9\right\}. \tag{13}$$

The Rényi family of generalized spectral dimensions is

$$D_q = \exp\left(\frac{\log\left(\sum_{k=1}^{R} p_k^q\right)}{1-q}\right), \qquad q > 0, \quad q \neq 1, \tag{14}$$

with $D_1 = r_{\text{eff}}$ by continuity and $D_2 = r_{\text{part}}$. Lower orders are more sensitive to the spectral tail, whereas higher orders emphasize dominant components. Comparing several orders therefore assesses whether the observed change depends on a particular spectral summary.

We further define the cumulative spectral mass and normalized tail energy:

$$C_K = \sum_{k=1}^{K} p_k, \qquad T_K = 1 - C_K. \tag{15}$$

For sorted spectra padded with zeros to a common length, the conditions $C_K^{\mathrm{CPS}} \le C_K^{\mathrm{control}}$ at every nontrivial K, together with equal total mass, mean that the CPS spectrum is majorized by the control spectrum. Larger T_K indicates a greater fraction of sample-dependent directional variance outside the leading K-dimensional subspace.

When reporting relative tail-energy gains, we use

$$G_K = 100\left(\frac{T_K^{\mathrm{CPS}}}{T_K^{\mathrm{control}}} - 1\right)\%, \tag{16}$$

for $T_K^{\mathrm{control}} > 0$. These statistics characterize representation geometry; they do not independently measure the amount of task-relevant localization information.

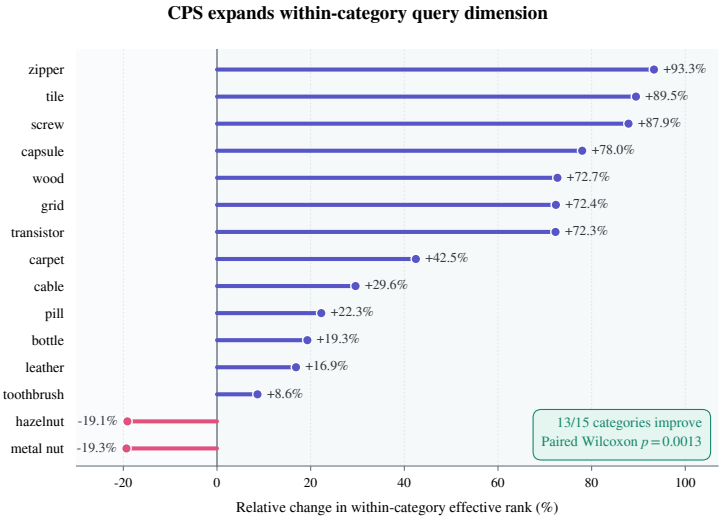


Figure 7: Category-level relative changes in the within-category effective rank of projected `[SEG]` queries on MVTec-AD after introducing CPS. Queries are extracted using a fixed inference instruction. Blue and orange bars indicate increases and decreases, respectively; categories are sorted by relative change. CPS increases effective rank in 13 of 15 categories, indicating a broad but non-universal increase in image-dependent query diversity.

A.2 Category-Level Analysis of Query Diversity

To examine whether the increase in query diversity is broadly shared across categories rather than driven by a few categories or between-category differences, we evaluate the within-category effective rank of projected `[SEG]` queries separately for each of the 15 MVTec-AD categories. All queries are extracted under the same fixed inference instruction, allowing us to assess image-dependent variation within each category.

As shown in Fig. 7, CPS increases within-category effective rank in 13/15 categories. The relative gains vary across categories, with the largest increases observed for zipper (93.3%), tile (89.5%), and screw (87.9%). Positive changes also occur in categories with more modest gains, such as leather (16.9%) and toothbrush (8.6%), indicating that the effect extends across a broad range of categories. Across all 15 categories, the mean paired absolute gain in effective rank is 3.91, with a category-bootstrap 95% confidence interval of $[1.92, 5.68]$. A one-sided paired Wilcoxon signed-rank test further supports an overall increase ($p = 0.00131$).

Hazelnut and metal nut are the two exceptions, with relative decreases of 19.1% and 19.3%, respectively. These exceptions indicate that the effect is broad but not universal, and that its magnitude depends on the category. Taken together, the category-level results are consistent with the motivation for CPS: complementary supervision during training mitigates the concentration of sample-

Table 7: Ablation of PAD and CPS with Qwen3.5-2B. Best results in **bold**.

PAD	CPS	AUROC	F1-max	AP
×	×	92.6	49.2	47.6
✓	×	**93.0**	49.6	48.7
×	✓	92.7	49.5	48.9
✓	✓	92.8	**50.4**	**49.5**

Table 8: Effect of teacher–student alignment layer selection on MVTec-AD. The shaded row denotes the default setting. Best results are shown in **bold**.

Aligned Layers	AUROC	F1-max	AP
6	92.3	50.7	48.8
12	92.1	50.9	48.0
18	92.6	51.1	49.0
24	**92.9**	**51.3**	**49.9**
6, 12, 18, 24	92.7	**51.3**	49.1

dependent query variation, even when inference uses a single fixed instruction. This analysis provides evidence of more diverse query representations, rather than directly establishing improved segmentation performance in every category.

A.3 Ablation with Qwen3.5-2B

To further examine the effects of PAD and CPS across MLLM backbones, Table 7 repeats the component ablation with Qwen3.5-2B. PAD alone achieves the highest AUROC of 93.0, improving over the baseline by 0.4 points, while CPS alone yields a smaller gain of 0.1 points. Combining PAD and CPS gives the highest F1-max and AP, reaching 50.4 and 49.5, respectively. Compared with the baseline, the combined setting improves AUROC/F1-max/AP by 0.2/1.2/1.9 points, indicating that PAD and CPS remain effective when applied to a different MLLM backbone, with their gains more evident in foreground localization metrics than in overall pixel-level ranking.

A.4 Effect of Alignment Layer Selection

We investigate how the depth of teacher–student affinity alignment affects PAD by comparing individual layers 6, 12, 18, and 24, as well as joint alignment across all four layers. All other settings remain unchanged. This comparison assesses whether supervision at multiple depths offers benefits beyond final-layer alignment.

As shown in Table 8, F1-max increases with depth among the single-layer settings, although AUROC and AP do not improve monotonically. Aligning layer 24 achieves the best overall results, with AUROC/F1-max/AP of 92.9/51.3/49.9. Joint alignment across all four layers matches its F1-max but reduces AUROC and AP by 0.2 and 0.8 points, respectively. Thus, incorporating shallower-layer constraints does not provide additional gains in this setting. These results support our default choice of final-layer alignment, which also avoids computing affinity losses at multiple depths.

A.5 Effect of Teacher Backbone Selection

We examine how the choice of affinity teacher affects PAD by comparing DINOv2-L, DINOv2-L with register tokens, DINOv3-L, and DINOv3-H+. The student backbone and training protocol remain unchanged. This comparison evaluates whether alternative teacher representations or a larger teacher improve the transfer of patch relationships.

As shown in Table 9, DINOv2-L achieves the best results on all three metrics, with AUROC/F1-max/AP of 92.9/51.3/49.9, outperforming DINOv3-L by 2.1/1.6/3.3 points. The DINOv2-L variant with register tokens yields lower scores by 0.9/0.7/2.0 points, indicating that this variant provides no benefit for affinity transfer in the evaluated setting. Although DINOv3-H+ improves over DINOv3-

Table 9: Effect of the affinity teacher backbone on MVTec-AD. Params. denotes the teacher parameter count. The shaded row denotes the default setting. Best results are shown in **bold**.

Teacher Backbone	Params.	AUROC	F1-max	AP
DINOv2-L	300M	**92.9**	**51.3**	**49.9**
DINOv2-L w/ Reg.	300M	92.0	50.6	47.9
DINOv3-L	300M	90.8	49.7	46.6
DINOv3-H+	840M	91.3	50.6	47.9

L on all three metrics, it remains below DINOv2-L despite its larger parameter count. These results support DINOv2-L as our default teacher and show that increasing teacher scale alone does not guarantee better downstream segmentation. All teachers are used only during training and introduce no additional inference cost.